AD-784 097

THE EFFECT OF DATA SOURCE RELIABILITY ON INTUITIVE INFERENCE

Edgar M. Johnson

Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia

July 1974

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Poit Royal Road, Springfield Va. 22151

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

J. E. UHLANER Technical Director	R. A. ROOTH COL, GS Commander

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 1300 Wilson Boulevard, Arlington, Virginia 22209.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
Technical Paper 251

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER
AD-784097

4. TITLE (and Subtitle)
THE EFFECT OF DATA SOURCE RELIABILITY ON INTUITIVE INFERENCE

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Edgar M. Johnson

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
U. S. Army Institute for the Behavioral and Social Sciences, 1300 Wilson Boulevard, Arlington, VA

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
20162101A754

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Intelligence Center and School, Ft Huachuca, AZ; U. S. Army Combined Arms Combat Development Activity, Ft Leavenworth, KS

12. REPORT DATE
July 1974

13. NUMBER OF PAGES
35

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Decision making
Bayes' Theorem
Reliability
Information processing
Subjective probability
Multistage inference

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

As part of a continuing effort to provide research information which will facilitate improved intelligence information processing, the Intelligence Information Processing Program undertook a study of man's processing and utilization of unreliable data. Reports from data sources of given reliability and diagnosticity were presented to 22 subjects in a series of two hypothesis decision problems. On each problem, each subject indicated the most likely of the two hypotheses and the subjective odds favoring that hypothesis. Subjective odds varied as a function of the data diagnosticity

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

and source reliability. However, the subjects generally failed to extract as much certainty as possible from the data--subjective odds were generally conservative with respect to odds computed by a normative Bayesian model. However, in most cases, as reliability decreased, subjective odds increased relative to Bayesian odds until they were generally greater than Bayesian odds at the lowest level of reliability. Subjects' protocols and data analyses indicated that subjects were using non-optimal inference strategies in which reliability was incorporated as a multiplicative weighting factor. If the diagnostic impact of the data ''if it were true'' is correctly evaluated, this strategy leads to increasingly inaccurate subjective odds as reliability decreases and the data diagnosticity increases.

Technical Paper 251

AD

THE EFFECT OF DATA SOURCE RELIABILITY ON INTUITIVE INFERENCE

Edgar M. Johnson

Robert S. Andrews, Supervisory Project Director

SYSTEMS INTEGRATION & COMMAND/CONTROL TECHNICAL AREA
Cecil D. Johnson, Chief

Approved By:
J. E. Uhlaner
TECHNICAL DIRECTOR

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

Office, Deputy Chief of Staff for Personnel
Department of the Army
1300 Wilson Boulevard, Arlington, Virginia 22209

July 1974

Army Project Number
20162101A754

Information Processing System a - 15

IV

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Papers are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The Intelligence Systems Work Unit within the U. S. Army Research Institute for the Behavioral and Social Sciences (ARI) is concerned with the functions of human information analysis, processing, aggregation, integration and consequent product utilization in intelligence systems. One of the major objectives is to provide research findings by which performance of these functions can be enhanced. One resulting requirement is to determine how human capabilities can be utilized to enable the intelligence information processing system to function with increased effectiveness. The entire research effort is responsive to requirements of RDTE Project 20162101A754, "Intelligence Information Processing," FY 1974 Work Program and to special requirements of the U. S. Army Intelligence Center and School.

The U. S. Army currently has under development intelligence information processing systems designed to maximize combat effectiveness by optimal utilization of human capabilities augmented by computer support. The present publication describes one effort which provides data for more effectively evaluating man's capabilities and limitations in intelligence processing.

J. E. UHLANER,
Technical Director

THE EFFECT OF DATA SOURCE RELIABILITY ON INTUITIVE INFERENCE

BRIEF

Requirement:

In order to develop improved methods for processing unreliable data into intelligence, research must be conducted to understand better how man processes and utilizes unreliable data in making inferences.

Procedure:

Reports from data sources of given reliability and diagnosticity were presented to 22 subjects in a series of two-hypothesis decision problems. On each problem, each subject indicated the most likely of the two hypotheses and the subjective odds favoring that hypothesis.

Findings:

Given data reports of known diagnosticity from a source of known reliability, subjective odds reflect variations in data diagnosticity and source reliability. The subjects generally failed to extract as much certainty as possible from the data--subjective odds were generally conservative with respect to odds computed by a normative Bayesian model. However, in most cases, as reliability decreased, subjective odds increased relative to Bayesian odds until they were generally greater than Bayesian odds at the lowest level of reliability.

Subjects' protocols and data analyses indicated that subjects were using non-optimal inference strategies in which reliability was incorporated as a multiplicative weighting factor. This strategy leads to increasingly inaccurate responses as reliability decreases and data diagnosticity increases, if the diagnostic impact of the data "if it were true" is correctly evaluated.

Utilization of Findings:

A large proportion of the data available to the intelligence system is of less than perfect reliability. The findings of the present study suggest several techniques for improving intuitive inferences based on unreliable data. Further research is required to test the utility of these techniques and to develop operational methods for improving intuitive inference.

THE EFFECT OF DATA SOURCE RELIABILITY ON INTUITIVE INFERENCE

CONTENTS

TABLES

FIGURES

THE EFFECT OF DATA SOURCE RELIABILITY ON INTUITIVE INFERENCE

INTRODUCTION

The introduction of tactical data systems for the Army in the field will substantially increase the amount and variety of data channeled into the intelligence system. A large proportion of these data will be ill-behaved--unreliable, dependent, redundant and/or of low resolution or completeness. In a manual intelligence system, such data tend to be filtered out by analysts who often discard low quality or questionable data because of time pressures. Although often discarded, such data could contribute substantially to the production of intelligence if improved methods and techniques of processing information can be developed. Tactical computers will afford the intelligence system an opportunity to systematically and logically incorporate ill-behaved data into the production of intelligence.

Thus, two broad questions are brought into focus: "What is the man in the system doing with the information available to him?" and "What should he be doing with it?" The first question raises a psychological issue which revolves around understanding how man processes and uses information. The second question is more practical and involves the development of aids and methods to enable more efficient and effective information processing. However, the problem of developing techniques for enhancing human performance in processing ill-behaved data requires that we first understand how man processes and uses such data.

The present study of ill-behaved data examines the ability of man to consider the reliability of a data source and the strategies he uses to process unreliable data when intuitive probabilistic inferences are required.

BACKGROUND

Source reliability can be viewed as a parameter having a direct effect upon the diagnosticity or impact of data. Previous research on probabilistic inference has shown that subjects are sensitive to changes in parameters that affect the diagnosticity of data but not necessarily in an optimum manner[1]. Although exceptions have been observed[2], the general finding,

[1] Peterson, C. R., and L. R. Beach. Man as an intuitive statistician. Psychological Bulletin, 1967, 68, 29-46.

[2] Schum, D. A. Inferences on the basis of conditionally non-independent data. Journal of Experimental Psychology, 1966, 72, 401-409.

termed conservatism, is that subjects respond as if the data were less diagnostic than they really are; that is, they fail to extract as much certainty as possible from the data[3]. Typically, increased data diagnosticity brings increased conservatism.

Often the processing or inference task is formally analogous to a problem in statistical inference, where items of evidence or data are used to determine the relative likelihood of alternative hypotheses. An optimal strategy for processing data in these tasks is Bayes' theorem, one form of which is:

$$P(H_i|D) = \frac{P(D|H_i)\ P(H_i)}{\Sigma_i P(D|H_i)\ P(H_i)} \qquad (1)$$

where $P(H_i)$ is the prior probability of a particular hypothesis; $P(D|H_i)$ is the probability of the occurrence of a particular item of data conditional upon the truth of a particular hypothesis; and $P(H_i|D)$ is the posterior probability of a particular hypothesis conditional upon the occurrence of a particular datum. Expressed in this way, the estimation of posterior probability is seen to involve two processes: first, the determination of the diagnostic impact of each datum ($P(D|H_i)$); and second, calculation of the posterior probability estimate ($P(H_i|D)$) on the basis of the observed data.

In inferring posterior probabilities from observations of data, subjects have been found to use a variety of non-Bayesian strategies[4]. Subjects may either revise a posterior probability by a constant regardless of the prior probability of the hypothesis or the diagnosticity of the data[5]; or they may base their responses on the similarity of the sample data to whatever representative feature of the hypothesis seems most relevant[6] or they may match their probabilities to the observed sample proportions[7]. Simon[8] suggests that, although non-optimal relative

[3] Slovic, P., and S. Lichtenstein. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

[4] Ibid.

[5] Pitz, G. F., L. Downing, and H. Reinhold. Sequential effects in the revision of subjective probabilities. Canadian Journal of Psychology, 1967, 21, 381-393.

[6] Dale, H. C. A. Weighing evidence: An attempt to assess the efficiency of the human operator. Ergonomics, 1968, 11, 215-230.

[7] Shanteau, J. C. An additive decision-making model for sequential estimation and inference judgments. Journal of Experimental Psychology, 1970, 85, 181-191.

[8] Simon, H. A. Models of man. New York: Wiley, 1957.

to Bayes' theorem, such strategies are rational. That is, in making inferences, man first cognitively constructs a simplified model of the real situation in order to deal with it. His behavior is consistent with respect to this model even though this behavior is not even approximately optimal with respect to the real world. This principle of bounded rationality suggests that as inference tasks become more complex (multistage), man will apply additional strategies for processing information which minimize cognitive complexity[9,10].

Data reliability can be incorporated into the Bayesian framework as another stage in the inference process. First, we must differentiate between the actual occurrence of a datum (D) and the report of its occurrence (D*). Assuming that the report of an event is not contingent upon which hypothesis is true, the conditional relationship between the the data and the hypothesis ($P(D|H_i)$) can be decomposed into[11,12]:

$$P(D^*|H_i) = P(D^*|D)P(D|H_i)+P(D^*|\bar{D})P(\bar{D}|H_i) \qquad (2)$$

where $P(D^*/D)$ is the probability of a report of some datum conditional upon the actual occurrence of that particular datum; $P(\bar{D}^*/D$ is the probability of a report of_some datum conditional upon the actual occurrence of any other datum; $P(\bar{D}/H_i)$is the probability of the occurrence of any other datum conditional upon the truth of a particular hypothesis; and $P(D/H_i)$ is as defined previously. Note that $P(\bar{D}/H_i)$ equals $1-P(D/H_i)$. Expressed in this way, the determination of the diagnostic impact of a report of some datum involves two processes, given a determination of source reliability ($P(D^*/D)$): first, determination of the diagnostic impact of the reported datum ($P(D/H_i)$) and the diagnostic impact of other data not reported ($P(\bar{D}/H_i)$); and second, calculation of the diagnostic impact of the report ($P(\bar{D}^*/H)$) on the basis of its reliability.

[9] Hormann, A. M. A man-machine synergistic approach to planning and creative problem solving: Part I. International Journal of Man-Machine Studies, 1971, 3, 167-184.

[10] Slovic, P. From Shakespeare to Simon: Speculations--and some evidence--about man's ability to process information. Oregon Research Institute, Eugene, Ore., Research Monograph, Vol. 12 No. 12, April 1972.

[11] Cavanagh, R. C., E. M. Johnson and R. L. Spooner. Multistage Bayesian inference systems. IEEE Transactions on Systems ,Man, and Cybernetics. in press.

[12] Schum, D. A., and W. M. DuCharm. Comments on the relationship between the impact and the reliability of evidence. Organizational Behavior and Human Performance, 1971, 6, 111-131.

The network of Figure 1 can be used to illustrate the flow of information when reliability is incorporated in a Bayesian model. Given the report of an event, a subject must first revise his opinion concerning which event actually occurred, before revising his opinion concerning the truth of an hypothesis. For example, if report D_1* is received, $P(D_1*|H_1)$ can be found by summing the products of the individual path segments from D_1 to H_1. However, the principle of bounded rationality suggests that in making such inferences man is likely to use a simpler, heuristic strategy. One such strategy might be called the "as if" approach, in which the subject treats the data as if it were perfectly reliable. That is, in processing unreliable data, the first stage of the inference process would be totally ignored; the diagnosticity of an event reported with a given reliability would have the same diagnosticity as the event itself.

A somewhat more complex heuristic strategy might be called the "best guess" approach. In this strategy, the subject tends to ignore the implications of less-likely data states in the transition from one stage to the next during the inference process and concentrates on the most likely data state[13]. The subject first evaluates the impact of the data as if it were perfectly reliable and then "shades" his estimate to reflect the reliability information. The subject may even construct a simplified model of the data network (Figure 1) and develop strategies similar to those described previously, in order to incorporate the impact of reliability information on the inference process.

Prior research, in which reliability was varied, indicates that subjects tend to overestimate the diagnostic impact of data reported with less than perfect reliability. In two experiments, Schum, DuCharme and DePitts[14] manipulated data reliability by varying subjects' observational uncertainty of tachistoscopically presented data. In both experiments, the excessiveness of subjects' posterior estimates was directly related to the diagnostic impact of the data. The experimenters noted that it was apparently not obvious to subjects that for a fixed reduction in reliability, the diagnostic impact of events with large inferential impact should be degraded more drastically than the diagnostic impact of events with lower inferential impact.

[13] Steiger, J. H., and C. F. Gettys. Best guess errors in multistage inference. Journal of Experimental Psychology, 1972, 92, 1-7.

[14] Schum, D. A., W. M. DuCharme and K. E. DePitts. Research on human multistage probabilistic inference processes. Rice University, Houston, Tex., Report No. 46-11, January 1971.

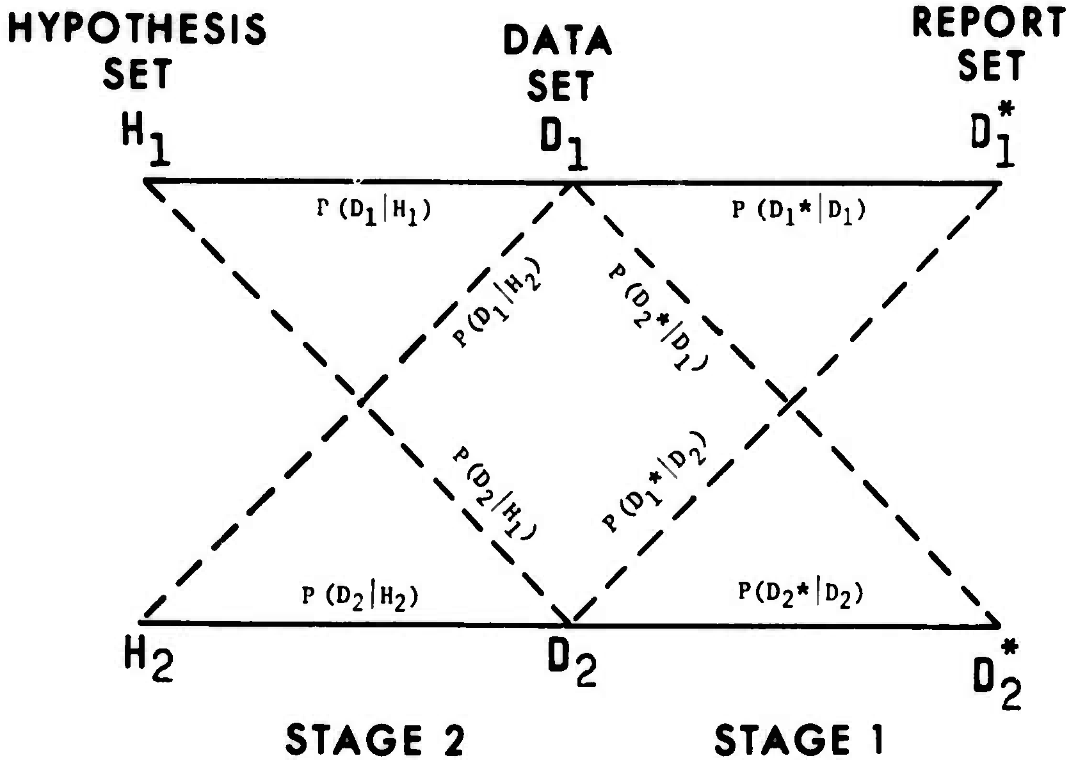

Figure 1. Conditional relationships in processing unreliable reports.

Snapper and Fryback[15] supported these results using a binary discrimination task with data from a source of known reliability. They suggested that subjects were using a non-optimal multiplicative model rather than the optimal Bayesian model. However, their analysis did not attempt to assess the fit of either a multiplicative model or of any alternative models.

Although prior research indicates that the optimal rule for processing unreliable information is not readily apparent to subjects, the actual strategies used and relative performance of subjects are unclear and only a limited range of experimental conditions have been used.

PURPOSE

The present study was designed to determine the relative performance of subjects and explore strategies used for processing information from less than perfectly reliable sources. The conditions were selected to cover a wide range of reliabilities and data diagnostic impact. A two-alternative decision task was used to provide a simple, easily understood task which required no prior training. The specific objectives were as follows:

1. To compare subjects' performance to a normative Bayesian model in weighting the diagnostic impact of a single datum or a sample of data from a source of known reliability.

2. To investigate subjects' strategies for incorporating information on the reliability of a data source in the inference process.

METHOD

Subjects

Twenty-two enlisted men, who had recently completed training as image interpreters, served as subjects. All had scored above 110 on the Army's General Technical aptitude area test.

[15] Snapper, K. J., and D. G. Fryback. Inferences based on unreliable reports. Journal of Experimental Psychology, 1971, 87, 401-404.

Inference Task

The classic urns and balls problem was used to provide a simple, easily understood inference problem. The task is well defined and consists of estimating from which of two urns a sample is most likely to have been drawn. One of the urns, say A, contains a certain percentage of red balls (P_1) and of blue balls ($1 - P_1$). This is the predominately red urn. The second urn, B, is the predominately blue urn containing P_0 red balls and ($1 - P_0$) blue balls. The problem was further simplified: first, by making the two urns symmetrical, the percentage of red balls in the predominately red urn was equal to the percentage of blue balls in the predominatly blue urn, that is, $P_1 = (1 - P_0)$; second, by making the prior probability of selecting an urn equally likely, that is, $P(A) = P(B) = .50$. Since the data are drawn with replacement from two categories (red and blue balls), samples follow a binomial probability distribution.

To estimate the most likely urn, the subject must consider three sets of information:

a. The composition of the urns, that is, the proportion of red and blue balls in both A and B. This is referred to as data generator diagnosticity, with diagnosticity increasing when the difference $P_1 - P_0$ increases;

b. The sample characteristics, or the total number of balls in the sample and the number of balls of one color in the sample, referred to as sample diagnosticity. Diagnosticity increases with increases in the difference between the number of balls of each color in the sample; and

c. The reliability of the source reporting the sample.

The diagnostic value of an event (a sample of balls from one of the two urns) is a function of all three types of information.

On each problem subjects first indicated which of the two urns they considered to be most likely. They then indicated their subjective odds in favor of the sample being from the most likely urn; that is, how many times more likely they considered the sample to be from the most likely urn than from the least likely urn. All numerical estimates were in the form of X:1, where $X \geq 1$.

Independent Variables

Sample Size. Two sample sizes were used: 1 datum and 5 data. In the 5-data sample condition, the sample was based on five independent draws of one ball each, with replacement, from the urn. The individual draws were not reported to the subject, only the cumulative results of the five independent draws.

Data Generator Diagnosticity. A symmetric pair of urns defined the data generator; that is, the set of conditional probabilities by which the data was generated. These are the probabilistic rules which govern the occurrence of data when a specific hypothesis is true. Four different urn compositions were used. In terms of the number of red balls in urn A, or equivalently the number of blue balls in urn B, these were 95, 90, 85 or 75 out of 100 total balls in each urn. Equivalently, the likelihood ratio for a single datum is 19:1, 9:1, 5.67:1 or 3:1, respectively, when expressed in terms of the more likely datum or color of ball from an urn.

Sample Diagnosticity. Sample diagnosticity refers to the relative diagnosticity of a particular data sample. Sample diagnosticity increases directly with the difference (d_i) between the number of balls of each color in the sample. In a sample of five data, d_i can be either one (d_1), three (d_3) or five (d_5). All three difference values occurred in the present experiment, the color of the most and least frequent balls in the sample being randomized.

Source Reliability. Reliability was defined as the percentage of reports from a source which were true. Reports were stated to come from one of five agents, X, Y, U, W and L of 60%, 70%, 80%, 90% and 100% reliability, respectively. The agents, except agent L, were pathological liars, and the occurrence of lies by an agent was independent of either the urn sampled or the color of the balls.

Experimental Materials

Three sets of problems were prepared for each subject--two sets of 16 one-datum sample problems and one set of 60 five-data sample problems. The problem sets were computer-generated and booklets were made using computer printout sheets with two problems per page (Figure 2). The 16 one-datum sample problems were composed using the four levels of data generator diagnosticity and the four levels of agent reliability excluding agent L. The 60 five-data sample problems were composed using the four levels of data generator diagnosticity, five levels of agent reliability and three levels of sample diagnosticity (Table 1). The 16 one-data sample problems are equivalent to the d_1 five-data sample problems. The four one-data sample problems with the 100% reliable agent were used as instructional examples. Note that for each sample size, all possible problems were used--either as test problems or as examples.

The following rules were used to order the problems within each problem set:

a. Urn A on the left, urn B on the right:

b. The predominately red urn equally often urn A and urn B;

c. The predominant color in the sample equally often red and blue;

d. The two problems on a page had different data generators, sample diagnosticity and source reliability.

```
25 58 59
 PAY CLOSE ATTENTION TO URN COMPOSITION
 AND AGENT RELIABILITY

          *         *        *         *
          *         *        *         *
          *         *        *         *
          * 75 RED  *        * 25 RED  *
          *         *        *         *
          * 25 BLUE *        * 75 BLUE *
          *         *        *         *
          *         *        *         *
          *         *        *         *
          ***********        ***********

            URN A              URN B

      AFTER 5 SAMPLES AGENT Z REPORTS 1 RED AND 4 BLUE
      AGENT Z IS 60 PERCENT RELIABLE

      THE MOST LIKELY URN IS  A  B    (CIRCLE ONE)
      ODDS FAVORING THIS URN ARE    TO 1

*********************************************

25 60 26
 PAY CLOSE ATTENTION TO URN COMPOSITION
 AND AGENT RELIABILITY

          *         *        *         *
          *         *        *         *
          *         *        *         *
          * 90 RED  *        * 10 RED  *
          *         *        *         *
          * 10 BLUE *        * 90 BLUE *
          *         *        *         *
          *         *        *         *
          *         *        *         *
          ***********        ***********

            URN A              URN B

      AFTER 5 SAMPLES AGENT Y REPORTS 4 RED AND 1 BLUE
      AGENT Y IS 70 PERCENT RELIABLE

      THE MOST LIKELY URN IS   A   B    (CIRCLE ONE)
      ODDS FAVORING THIS URN ARE     TO 1
```

Figure 2. Problem format

Table 1

LIKELIHOOD RATIOS OF EVENT IMPACT CONDITIONS

Data	Sample	Reliability				
Generator	Difference	1.	.9	.8	.7	.6
19:1	1	19.00	6.14	3.35	2.13	1.44
	3	6859.00	231.80	37.52	9.60	2.98
	5	2476099.00	8746.86	420.55	43.33	6.17
9:1	1	9.00	4.56	2.85	1.94	1.38
	3	729.00	94.54	23.06	7.31	2.63
	5	59049.00	1962.03	186.76	27.56	5.02
5.67:1	1	5.67	3.55	2.45	1.78	1.33
	3	181.96	44.57	14.68	5.62	2.33
	5	5843.03	560.22	87.96	17.76	4.09
3:1	1	3.00	2.33	1.86	1.50	1.22
	3	27.00	12.70	6.91	3.38	1.83
	5	243.00	69.16	22.09	7.59	2.73

In each problem set the predominant color in a sample was balanced for combinations involving sample diagnosticity, data generator diagnosticity and agent reliability in all possible pairs. This was not done for the combinations of all three factors. Each subject received a different random ordering of the problems on each problem set.

Experimental Design

The experiment was divided into two main designs corresponding to sample size. The one-datum sample problems constituted a three-factor completely within-subjects design. The factors were experience--before and after the five-data sample problems; source reliability--four levels of agent reliability; and data diagnosticity--four levels of data generator diagnosticity. The five-data sample problems also constituted a three-factor completely within-subjects design: source reliability--five levels of agent reliability; data diagnosticity--four levels of data generator diagnosticity; and three levels of sample diagnosticity.

Procedure

The subjects were run in two groups--one of seven subjects and one of 15 subjects--corresponding to the number of students available from two consecutive image interpreter classes. All subjects served in each condition. Each session was broken into four phases: instructions, one-datum sample problems, five-data sample problems, and a second set of one-datum sample problems.

Prior to the first set of one-datum sample problems, subjects were briefed concerning the general nature of the experiment (decision making) and told not to discuss the experiment among themselves until its conclusion. Each subject was then given a set of four sample problems of one-datum reported by agent L (100% reliable), one problem for each of four different urn compositions. Before working the sample problems, the subjects were instructed to:

> Assume that I take the two urns shown into the next room. I will choose one of the two urns by flipping a fair coin: heads I'll choose urn A, tails I'll choose urn B. After choosing an urn I'll choose a ball, without looking, from the urn. Once the ball is drawn from the urn I'll give it to an agent who will report the color of the ball. Given the color of the ball that was drawn, I want you to first choose the most likely urn, A or B, by circling the appropriate letter. Secondly, I want you to write the odds favoring this urn. That is, how many times more likely is it that the ball was taken from the most likely urn than from the least likely urn. Note that this is always a number greater than one.

After the subjects worked through the four problems at their own pace, the experimenter answered any questions and insured that all subjects understood the correct odds for each problem. Subjects were then given instructions on agent reliability:

> In the problems you just finished, all of the reports were from agent L who was stated to be 100% reliable. He always reports the correct color of the ball drawn from the urn. In this next set of problems, there are four new agents--X, Y, U and W. All of these agents are liars. They do not always report the correct color of the ball which was drawn from the urn. Sometimes, they will report that the opposite color was drawn. As they don't know from which urn the ball was drawn, their lies about the color of a ball do not depend on which urn the ball was drawn. However, they don't lie all the time. For example, agent X is 60% reliable and lies only 40% of the time, and agent W is 90% reliable and lies 10% of the time. In this series of problems, the color of the ball drawn will be reported by one of the agents who lies.

An assistant demonstrated the concept of the "liar" by lying on one of four independent samples from an urn.

The experimenter answered any questions and distributed test booklets containing the first set of one-datum problems. Following completion of the first set of one-datum sample problems, the test booklets were collected. Subjects were then instructed that the next problem set would contain reports based on five independent draws with replacement. Thus, the report they received was based on the cumulative result of five independent reports from the liar.

The experiment answered any questions and distributed the new test booklets. When all of the subjects had finished the five-data sample problems and turned in the test booklets, they were given a five-minute break, before receiving the last set of one-datum sample problems.

After completing the last problem set, subjects filled out a questionnaire relating to the experiment and were asked to explain the method they had used for computing the odds favoring the most likely urn. After all subjects finished the questionnaire, the Bayesian solution was explained and any questions concerning the experiment were answered by the experimenter.

A session lasted approximately two hours and subjects were allowed to work at their own pace within each problem set. Subjects were permitted individual breaks during the session in addition to the scheduled five-minute break.

RESULTS

A subjects' choice of the most likely urn on each problem represents a dichotomous score. However, it yielded little information concerning decision strategies; out of the 1,320 problems performed by the 22 subjects, there was only one instance of an error or problem on which the least likely urn was chosen. The problem--a report of three red and two blue from data generators with 19:1 odds by agent Y (70% reliable)--was annotated by the subject saying that the agent was lying. The second part of the response, subjects' odds, constitutes the primary data. All of the following analyses are based on 21 subjects; one subject was deleted for giving odds responses of less than one.

The first question to be addressed is whether subjective odds were sensitive to manipulations of the independent variables. Figure 3 shows subjects' mean odds as a function of data generator and reliability for each level of sample diagnosticity. A three-way analysis of variance--Data generator x Sample diagnosticity x Reliability--was performed on subjects' log odds. (A log transform was used to stabilize the within-subject variance which was nonhomoscedastic.) Reliability was a significant main effect, $F(4,80) = 48.27$, $p < .01$, and each level of reliability was significantly different from every other level of reliability using Tukey's Honestly Significant Difference Test ($p < .05$). There were significant main effects of data generator, $F(3,60) = 75.18$, $p < .01$, and of sample diagnosticity, $F(2,40) = 5.56$, $p < .01$; and the interaction between these latter two factors was also significant, $F(6,120) = 2.15$, $p < .05$. Tests of the simple main effect of sample diagnosticity were significant, ($p < .01$), except for the least diagnostic data generator of 3:1. Comparisons among levels of sample diagnosticity for each data generator indicated that d_1 was significantly different ($p < .05$) from d_3 and d_5, but d_3 and d_5 were not significantly different from each other (Figure 3). This interaction is evident in Figure 3 as an increased spread between the plots for each generator as reliability increases. No other effects were significant. This analysis indicates that subjects were sensitive to the independent variables; decreasing reliability, data diagnosticity, or sample diagnosticity led subjects to decrease their odds.

Sample Size

In generating the one-datum sample problem booklets, a computer programming error resulted in the first group of 15 subjects receiving a non-orthogonal problem set in which some problems were missing and others appeared more than once. This programming error was corrected and the second group of seven subjects received orthogonal problem sets. There were no apparent differences in the subjective odds estimates of subjects in the second group between the two sets of 16 one-datum sample problems or between subjects in the two groups. Therefore, subjects from the two

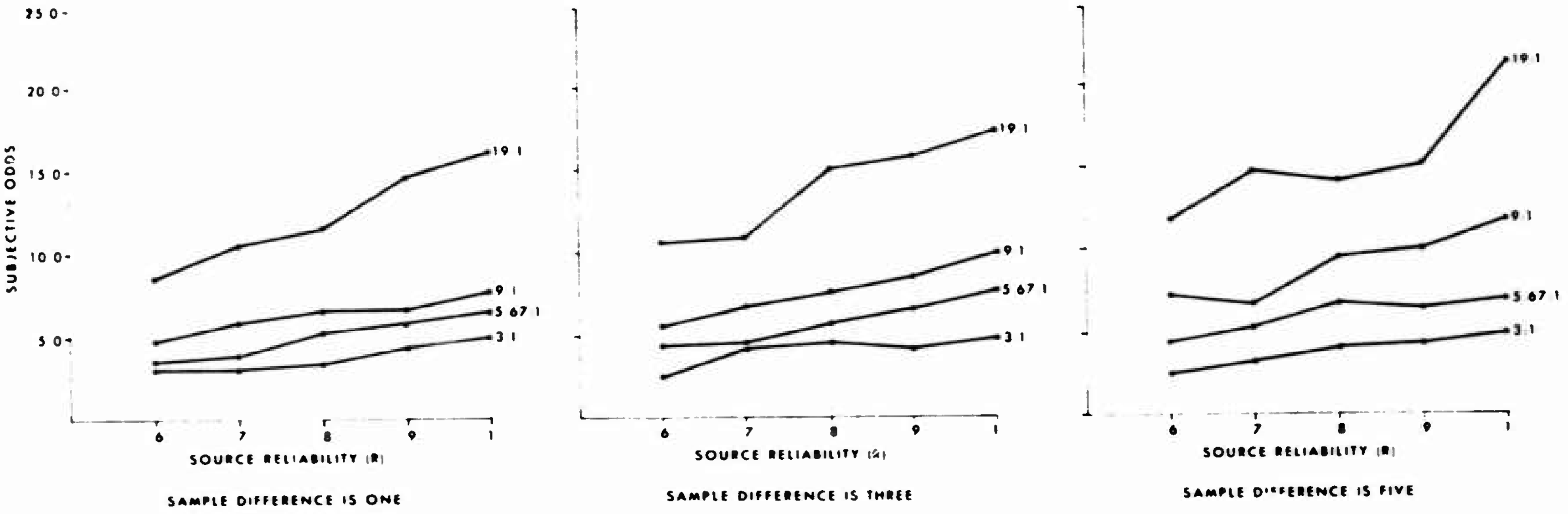

Figure 3 Mean sucjective odds for each data generator.

groups were pooled to compare responses on the 16 one-datum sample problems with the 16 isomorphic five-data sample problems. The isomorphic five-data sample problems are those in which sample diagnosticity, or the difference in the number of balls of each color, was one and the source reliability was not 100%.

Eight subjects in the first group received all of the 16 one-datum sample problems at least once. The mean response was used whenever a problem occurred more than once in any subject's two-problem sets. These data were then pooled with the mean responses for the seven subjects in the second group. The data from these 15 subjects were used to analyze the effect of sample size on subjective odds in a three-way analysis of variance of Sample size x Data generator x Source reliability. There was no significant main effect or interaction involving sample size. Apparently subjects did not process isomorphic information differently for the two sample sizes. Thus, further analyses are based only on responses from the five-data sample problem set.

Decision Performance

The analyses thus far have indicated that subjects' odds were influenced by data generator, sample diagnosticity, and reliability, but not sample size. However, these analyses give no clue to the quality of their decisions or the strategy used.

A useful index of a subject's efficiency as an information processor is the difference measure, Δ: the subject's odds minus the criterion odds[16]. When the criterion is the Bayesian odds, the index, Δ_B, indicates in log units the ratio of subjective odds to the corresponding Bayesian odds. A negative value of the index indicates conservatism and a positive value indicates extremism, while a zero value indicates optimal performance. A conservative response represents an error of extracting less certainty than available in the data, whereas, an extreme response represents an error of extracting more certainty than available. Figure 4 shows mean Δ_B as a function of data generator and reliability for each level of d. These plots resemble those from other inference tasks with reliable sources. The more diagnostic the sample, the less optimal the subjective odds. As the sample becomes less diagnostic, subjects' responses come closer to being optimal, and finally with very undiagnostic data the responses are extreme. As reliability decreased, subjects' odds increased relative to Bayesian odds. Subjects were not only influenced by reliability, but in fact became more Bayesian as reliability decreased, except for d_1, in which case subjects' responses became extreme.

[16] Shum, 1966.

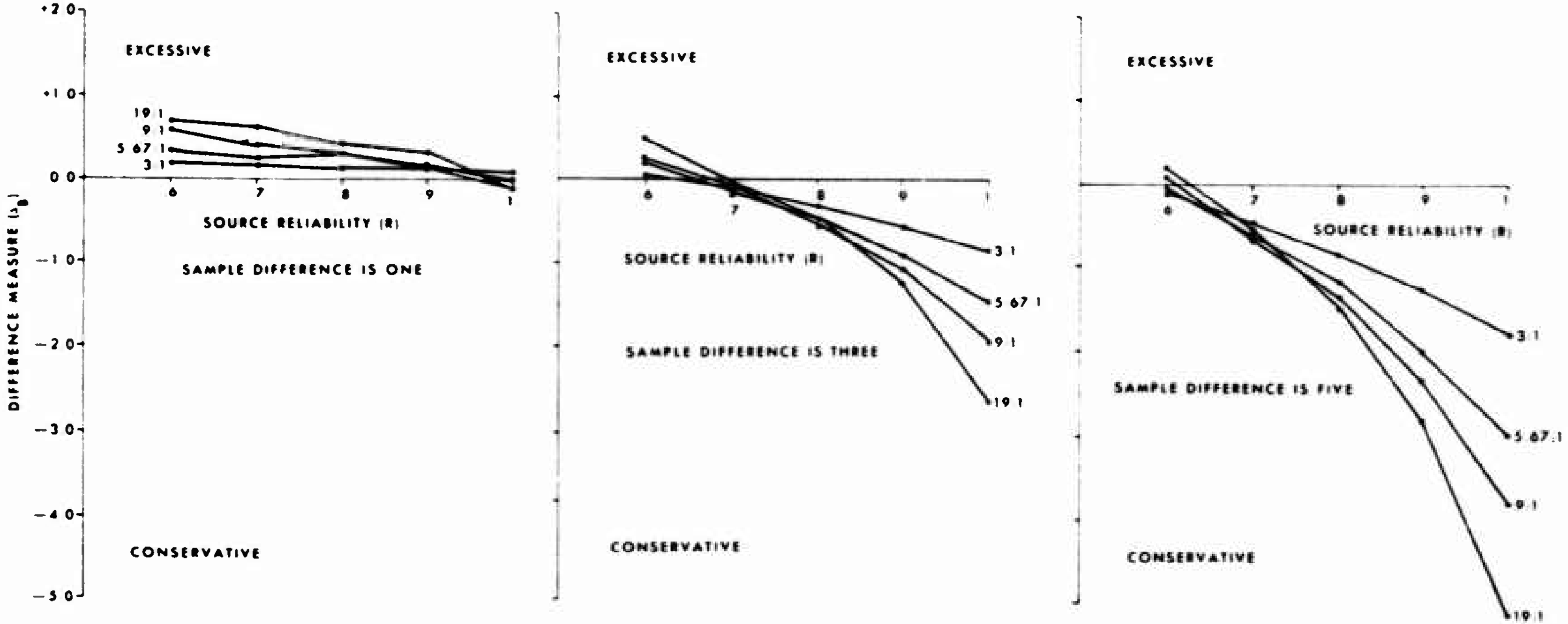

Figure 4. Mean difference measure (Δ_B) of subjective odds with Bayesian odds.

The increase in Δ_B with decreasing reliability appears to result from subjects weighting reliability similarly for all combinations of data generator and sample diagnosticity. It is clear from Figure 4 that the addition of another level of processing required to incorporate reliability into a Bayesian model does not lead to greater conservatism. However, we can easily question the applicability of a Bayesian model as an analogue of the subjective inference process in this task.

Decision Strategy

On the post-experiment questionnaire, subjects were asked to explain the strategy they used in the task. In deciding which urn was most likely, all 21 subjects reported always choosing the urn with the larger proportion of balls of the predominant color in the sample. In assigning odds, 20 of the subjects reported a strategy of multiplying the odds of drawing one ball of the predominant color in the sample from the most likely urn by the difference in the number of balls; and then multiplying the product by the agent's reliability transformed to a probability. Eight of the 21 subjects reported shading this value when they thought the agent was lying.

Subjects' reported decision strategy and the relatively constant slope of the graphs of subjective odds as a function of reliability suggest that subjects were using a simple multiplicative rather than a Bayesian inference rule. In a multiplicative strategy, odds are obtained by multiplying the odds in favor of drawing one ball of the predominant color in the sample from the most likely urn by the difference between the number of blue and red balls in the sample, and then multiplying the product by the reliability of the source.

The fit of this rule to subjects' odds was investigated using the difference measure with the simple multiplicative rule as the criterion, Δ_{SM}. Mean Δ_{SM} as a function of data generator and reliability at each level of d is shown in Figure 5. Negative values of Δ_{SM} indicate excessive estimates where the multiplicative rule overestimates subjects' odds and positive values indicate an underestimate of subjects' odds. The relatively flat slopes of the graphs of Δ_{SM} indicate that this rule predicts subjects' use of source reliability information. However, the spread between data generators at each level of d and the differences between the graphs at each level of d indicate that subjects' use of sample diagnosticity information is not by a simple multiplicative strategy. A three-way analysis of variance of Δ_{SM}--Data generator x Sample diagnosticity x Reliability--had only two significant effects: data generator, $F(3,60) = 6.94$, $p < .01$, and d, $F(2,40) = 218.96$, $p < .01$.

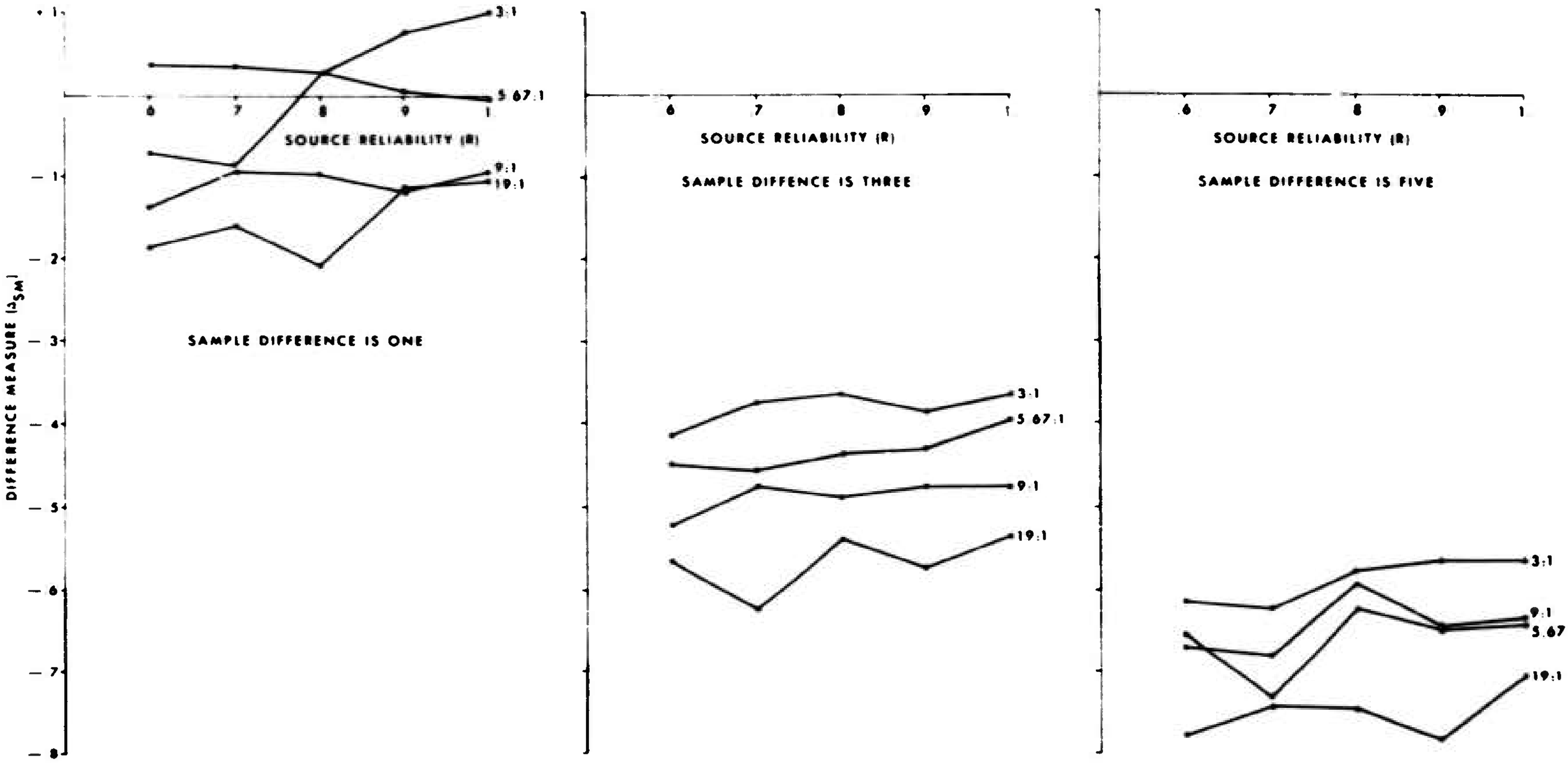

Figure 5. Mean difference measure (Δ_{SM}) of subjective odds with odds of simple multiplicative rule

The above results suggest that subjects are not using a simple multiplicative strategy to combine information about the data generator and sample diagnosticity but do include source reliability as a multiplicative factor. This implies a slightly different subjective strategy which may be termed the derived multiplicative rule. In this strategy, the odds for a problem with a given source reliability (not equal to 100%), data generator, and sample diagnosticity are obtained by multiplying the reliability by the odds provided by the subject on the problem where the source was perfectly reliable (100%) and which had the same urn composition and sample diagnosticity. This rule assumes that the subject uses some unspecified intuitive method to combine information about the data generator and sample diagnosticity, and then includes source reliability as a multiplicative factor. Using a difference measure with this rule as the criterion, Δ_{DM}, the mean Δ_{DM} over all subjects and all conditions was -0.03. A three-way analysis of variance of Δ_{DM}--Data generator x Sample diagnosticity x Reliability--had no significant main effects or interactions. This rule slightly overestimates subjects' odds, but otherwise it is a good fit.

Another view of how well subjective odds match the odds that would be produced by employing each of these three strategies is obtained using a correlation analysis. Table 2 gives the product moment correlation coefficients between each subject's odds and the odds predicted by each rule, the Bayesian, the simple multiplicative, and the derived multiplicative. All 60 problems were used in computing the correlations for the Bayesian and simple multiplicative rules; the 48 problems that did not involve the 100% reliable source were used in the computations for the derived multiplicative rule. For nearly all subjects, the coefficients are ordered, in increasing size, Bayesian, simple multiplicative, and derived multiplicative, indicating an increasing match between subjective and predicted odds. The average coefficients and the percentage of variance-in-the-prediction accounted for by the rules were .31, 10.3% for the Bayesian; .65, 44.4% for the simple multiplicative; and .80, 67.0% for the derived multiplicative. In sum, the derived multiplicative rule was most successful in describing subjective performance.

The average correlations observed between the Bayesian odds and those obtained according to the simple multiplicative rule (.46) and between the Bayesian odds and those obtained according to the derived multiplicative rule (.36) raise the possibility that the correlations between subjective odds and those predicted by the multiplicative rules may be artifically inflated to the extent that the subjects were actually using the Bayesian rule. This possibility was tested by partialing out the variance due to the Bayesian odds: the resulting average partial correlations were .60 between subjective odds and simple multiplicative rule odds, and .78 between subjective odds and derived multiplicative rule odds. These high partial correlations provide further support that subjects followed versions of a multiplicative rule rather than the optimal, Bayesian rule.

Table 2

COEFFICIENTS OF CORRELATION BETWEEN SUBJECTIVE AND PREDICTED ODDS

Subject Number	Rule		
	Bayesian (N=60)	Simple Multiplicative (N=60)	Derived Multiplicative (N=48)
1	.17*	.55	.77
2	.36	.88	.93
3	.28	.70	.89
4	.30	.50	.49
5	.27	.52	.84
6	.24	.37	.53
7	.31	.72	.99
8	.30	.72	.97
9	.29	.32	.53
10	.26	.65	.88
11	.31	.72	.99
12	.42	.88	.77
13	.24	.71	.99
14	.28	.72	.99
15	.14*	.36	.56
16	.37	.92	.99
17	.40	.55	.59
18	.41	.58	.71
19	.31	.72	.94
20	.58	.77	.52
21	.23	.70	.88

*Not significant; all other correlations are significant at $p < .05$.

The number of semesters of college or the number of semesters of college mathematics had no discernible relationship with either attitudes toward, performance on, or the strategy used on the inference task.

DISCUSSION

The results of this research indicate that subjective odds reflect variations in all three independent variables--data generator, sample diagnosticity, and source reliability--which affected report diagnosticity. Subjects' protocols and data analyses indicate that subjects used a multiplicative strategy to incorporate reliability into the inference process. However, data generator diagnosticity and sample diagnosticity were apparently not combined multiplicatively (Table 2). This strategy reflects the two-stage structure of the inference problem. Subjects first estimated odds "as if" the report were true and then weighted these odds by multiplying the "as if" odds by the stated reliability. Snapper and Fryback[17] also found that subjects used a multiplicative rule to combine reliability information in the inference process. However, they also indicated that data generator diagnosticity combined multiplicatively. This difference in results is probably due to their use of only one-data sample problems. These results provide an explanation of results in an earlier study in which it was found that subjects overpaid for unreliable data in an information purchase task[18].

A multiplicative strategy is not a "best guess" strategy in the sense of a tendency to ignore the implications of less likely events in the transition from one stage to the next in the inference process. Although non-optimal, subjects apparently constructed a simplified, but rational model in Simon's[19] terms. In terms of the data network of Figure 1, subjects were ignoring the cross-over effects: the implication that events other than the event reported may have occurred.

In general, subjective posterior odds decreased with decreasing reliability, data generator diagnosticity and sample diagnosticity. When the source was perfectly reliable, subjective odds were generally conservative with respect to those computed by Bayes theorem. In most cases, however, as reliability decreased, subjective odds increased relative to Bayesian odds until they were generally greater than Bayesian odds at the lowest level of reliability ($P(D^*/D)=.6$). Thus, the added information processing required to incorporate reliability into the inference process not only did not lead to greater conservatism but led instead to more extreme subjective odds.

[17] Snapper and Fryback, 1971.

[18] Kanarick, A. F., J. Huntington and R. C. Petersen. Multi-source information acquisition with optional stopping. Human Factors, 1969, 11, 379-386.

[19] Simon, 1957.

Note that the error produced by a multiplicative decision strategy serves to partially offset the usual conservative bias in subjective odds. It is apparently not obvious to subjects that, for a fixed reduction in reliability, reports with high diagnosticity should be degraded more than reports with low diagnosticity. However, the high diagnostic impact of highly diagnostic data is also apparently not obvious. Thus, the phenomenon of conservatism is offset by a non-optimal decision strategy.

Subjective posterior odds were not different between the one-datum sample and the five-data sample problems at the same level of diagnosticity. This result differs from the earlier work of Vlek[20] and Pitz[21] on the effects of sample size. However, these two studies used larger sample sizes which may account for the differing results.

Subjects' use of a simplified cognitive model of the task suggests three approaches to improving inference performance with unreliable data. These approaches are based on increasing the complexity of an inference maker's processing model. First, subjects could be given instruction on the structure of multistage inference problems. Second, in evaluating a report from an unreliable source, subjects could be required to list the other events which may have occurred, but which were not reported. The effect of either approach might only be to convert a subject's multiplicative model into a "best guess" strategy. However, the increased awareness of the complexity of an optimal model for incorporating reliability information into the inference process should result in a net improvement in performance.

A third approach to performance enhancement is to couple the inference maker to computer-supported information processing and decision-making aids. Complex multistage problems could be analytically solved, and in addition, the sensitivity of inferences to imput parameters could be assessed and "constant reliability" contours could be calculated and displayed. This information presented to an intelligence analyst via a real-time display or summary table could be used as an on-line inference aid or incorporated into a training program[22,23].

[29] Vlek, C. The use of probabilistic information in decision making. Psychological Institute Report No. 009-65, University of Leiden, The Netherlands, 1965.

[21] Pitz, G. F. Sample size, likelihood, and confidence in decision. Psychonomic Science, 1967, 8, 257-258.

[22] Johnson, E. M, and S. M. Halpin. Preliminary evaluation of a multistage Bayesian inference system. In Proceedings of the 1973 International Conference on Cybernetics and Society (IEEE), 1972, 431-435.

[23] Hammond, K. R. Computer graphics as an aid to learning. Science, 1971, 172, 903-908.

A final point which should be noted concerning the present study and earlier studies[24,25] is that they were conducted in laboratory settings using relatively simple tasks. The findings in these studies should be validated in more realistic and complex tasks.

CONCLUSIONS

The present study provides some insights into the process of intuitive inference when the data is of less than perfect reliability. The subjects used rational but non-optimal information processing strategies which increase in error as reliability decreases and data diagnosticity increases. However, the phenomenon of conservatism serves to partially offset the error resulting from this strategy. The findings need to be validated in more complex decision environments, and indicate a requirement for research oriented toward improving human inference performance with ill-behaved data.

[24] Schum, DuCharme, and Pitts, 1971.

[25] Snapper and Fryback, 1971.

REFERENCES

Cavanagh, R. C., E. M. Johnson, and R. L. Spooner. Multistage Bayesian inference systems. IEEE Transactions on Systems, Man, and Cybernetics, in press.

Dale, H. C. A. Weighing evidence: An attempt to assess the efficiency of the human operator. Ergonomics, 1968, 11, 215-230.

Hammond, K. R. Computer graphics as an aid to learning. Science, 1971, 172, 903-908.

Hormann, A. M. A man-machine synergistic approach to planning and creative problem solving: Part I. International Journal of Man-Machine Studies, 1971, 3, 167-184.

Johnson, E. M. and S. M. Halpin. Preliminary evaluation of a multistage Bayesian inference system. In Proceedings of the 1973 International Conference on Cybernetics and Society (IEEE), 1972, 431-435.

Kanarick, A. F., J. Huntington, and R. C. Peterson. Multi-source information acquisition with optional stopping. Human Factors, 1969, 11, 379-386.

Peterson, C. R. and L. R. Beach. Man as an intuitive statistician. Psychological Bulletin, 1967, 68, 29-46.

Pitz, G. F. Sample size, likelihood, and confidence in decision. Psychonomic Science, 1967, 8, 257-258.

Pitz, G. F., L. Downing, and H. Reinhold. Sequential effects in the revision of subjective probabilities. Canadian Journal of Psychology, 1967, 21, 381-393.

Schum, D. A. Inferences on the basis of conditionally non-independent data. Journal of Experimental Psychology, 1966, 72, 401-409.

Schum, D. A., and W. M. DuCharme. Comments on the relationship between the impact and the reliability of evidence. Organizational Behavior and Human Performance, 1971, 6, 111-131.

Schum, D. A., W. M. DuCharme, and K. W. DePitts. Research on human multistage probabilistic inference processes. Rice University, Houston, Tex., Report No. 46-11, January 1971.

Shanteau, J. C. An additive decision-making model for sequential estimation and inference judgments. Journal of Experimental Psychology, 1970, 85, 181-191.

Simon, H. A. Models of man. New York: Wiley, 1957.

Slovic, P. From Shakespeare to Simon: Speculations--and some evidence--about man's ability to process information. Oregon Research Institute, Eugene, Ore., Research Monograph, Vol. 12, No. 12, April 1972.

Slovic, P., and S. Lichtenstein. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Snapper, K. J. and D. G. Fryback. Inferences based on unreliable reports. Journal of Experimental Psychology, 1971, 87, 401-404.

Steiger, J. H., and C. F. Gettys. Best guess errors in multistage inference. Journal of Experimental Psychology, 1972, 92, 1-7.

Vlek, C. The use of probabilistic information in decision making. Psychological Institute Report No. 009-65, University of Leiden, The Netherlands, 1965.